DOCUMENT RESUME

ED 389 752 TM 024 385

AUTHOR van der Linden, Wim J.; Luecht, Richard M.
TITLE An Optimization Model for Test Assembly To Match Observed-Score Distributions. Research Report 94-7.
INSTITUTION Twente Univ., Enschede (Netherlands). Faculty of Educational Science and Technology.
PUB DATE Nov 94
NOTE 29p.
AVAILABLE FROM Bibliotheek, Faculty of Educational Science and Technology, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Ability; Foreign Countries; Heuristics; *Item Response Theory; Linear Programming; Mathematics Tests; Models; Observation; *Scores; *Statistical Distributions; *Test Construction
IDENTIFIERS ACT Assessment; *Optimization

ABSTRACT

An optimization model is presented that allows test assemblers to control the shape of the observed-score distribution on a test for a population with a known ability distribution. An obvious application is for item response theory-based test assembly in programs where observed scores are reported and operational test forms are required to produce the same observed-score distributions as long as the population of examinees remains stable. The model belongs to the class of 0-1 linear programming models and constrains the characteristic function of the test. The model can be solved using the heuristic presented in Luecht and T. M. Hirsch (1992). An empirical example with item parameters from the ACT Assessment Program Mathematics Test illustrates the use of the model. (Contains 6 figures and 23 references.) (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 389 752

An Optimization Model for Test Assembly to Match Observed-Score Distributions

Research Report 94-7

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. NELISSEN

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Wim J. van der Linden

Richard M. Luecht

BEST COPY AVAILABLE

faculty of
EDUCATIONAL SCIENCE AND TECHNOLOGY

University of Twente

Department of Educational Measurement and Data Analysis

TM024386

An Optimization Model for Test Assembly to Match Observed-Score Distributions

Wim J. van der Linden
Richard M. Luecht

To appear in G. Engelhard & M. Wilson (Eds.), Objective measurement: Theory into practice (Volume 3). Norwood, NJ: Ablex.

An optimization model for test assembly to match observed-score distributions, Wim J. van der Linden and Richard M. Luecht - Enschede: University of Twente, Faculty of Educational Science and Technology, November, 1994 - 23 pages.

Abstract

An optimization model is presented that allows test assemblers to control the shape of the observed-score distribution on a test for a population with a known ability distribution. An obvious application is IRT-based test assembly in programs where observed scores are reported and operational test forms are required to produce the same observed-score distributions as long as the population of examinees remains stable. The model belongs to the class of 0-1 Linear Programming models and constrains the characteristic function of the test. The model can be solved using the heuristic presented in Luecht and Hirsch (1992). An empirical example with item parameters from the AAP Mathematics Test illustrates the use of the model.

An Optimization Model for Test Assembly to Match Observed-Score Distributions

A traditional objective in test assembly is to maximize the reliability of the test. From a classical test theory point of view, maximum reliability is an attractive feature of a test because tests with a high reliability are sensitive to the differences in true scores between the examinees in the population and have a low standard error of measurement. With the advent of item response theory (IRT), however, the objective of test assembly changed and it became possible to assemble tests to meet a targeted information function. It was Birnbaum (1968) who paved the way for this new objective, proposing an attractive two-stage test procedure based on item and test information functions. The first step in Birnbaum's procedure is to establish a target for the information function of the test. This requirement forces test assemblers to think about the intended use of the test scores and its translation into an optimal distribution of the information in the test scores along the ability scale. Once a target for the test information function is established, the test is assembled such that the sum of the item information functions matches the target for the test information function.

The objective addressed in this paper is new in that a model for test assembly from a calibrated item pool is presented to match a target for the *observed-score distribution* for a given population of examinees. This objective may seem unusual because it the shares the assumption of an IRT-calibrated item pool with an explicit interest in observed-score distributions; something usually associated with classical test theory. However, this new test assembly model perfectly reflects much

of modern testing practice, where IRT is increasingly used to produce high-quality tests (i.e., using IRT item parameters estimates to assemble the test or to pre-equate test forms) but test scores are still reported on an observed-score scale. This practice can be found, for example, in testing programs with observed-score scales established before IRT was introduced and where it is impossible to change reporting practices without upsetting the consumers.

In such testing programs, it is important to have control of the observed-score distribution. If changes in the testing program are introduced, the practical consequences of these changes could be minimized if test assembly would offer the possibility to explicitly control their effects on the observed-score distribution. Examples of changes in testing programs that may effect observed-score distributions are: (1) the introduction of new specifications for the item pool; (2) a change of item calibration procedures; and (3) item parameter drift. It should be noted that the intent here is not to control individual scores but only their distribution. This approach is applicable, for example, if some items with new specifications are added to the pool leading to minor changes in the relative abilities of examinees, whereas the same observed-score distribution represents the order between the abilities as adequately as the old pool.

Alternative Solutions

An attempt to control the observed-score distribution is also present in some procedures already in use in educational measurement. A few examples of such procedures are: equipercentile equating, item matching, and test assembly using target information functions.

In equipercentile equating, the cumulative distribution function of the observed scores on a new test form is equated to the same function of an old test

form. However, equipercentile equating can only take place after the new test form is administered. In addition, this method of equating obtains its results by distorting the observed-score scale of the new test form. A better solution would be to assemble all new test forms to automatically produce the required distribution of observed scores. Attempting to achieve the latter solution is a fundamental rationale for the procedure described here. It is, however, correct to view this new procedure as a variant of equipercentile *pre*-equating.

With item matching, a new test form can be matched item by item to an old test form. One method introduced to realize this objective is Gulliksen's (1950) Matched Random Subsets Method. Linear Programming (LP) models that implement Gulliksen's method are given in Amstrong and Jones (1992) and van der Linden and Boekkooi-Timminga (1988). If items are matched on the basis of estimates of parameters describing their marginal and joint distributions, for example, item p-values and covariances, then two test forms with perfect match are bound to produce identical observed-score distributions for the same population of examinees. However, methods of item matching may involve new and stringent constraints on a test assembly process in addition to all other constraints that are typically needed (e.g., the test content, the format of the items, the length of the item-related text, and the distribution of the keys across response alternatives). As a result, in practice, perfect matches may not be approached closely enough to produce satisfactory observed-score distributions. The model proposed in this paper is not restricted by any new constraints on the assembly process.

Finally, it is possible to assembly a test using target information functions. A popular definition of parallel tests in IRT is Samejima's (1977) which considers tests to be parallel if they have the same information function. However, unlike classical definitions of parallel tests, Samejima's definition does not guarantee

identical observed-score distributions. One reason is that in test assembly two different sets of item response functions may approach the same target information function. A more fundamental reason, however, is that a test information function only governs the (asymptotic) distribution of error in the ability estimates on the Θ-scale but not the distribution of the true scores for the test.

An Optimization Model

The approach in this chapter is to assemble a test using a target for the characteristic function rather than the information function of the test. This characteristic function is the transformation needed to transform the Θ-scale in the IRT model into the true-score scale underlying the test. The true-score scale is identical to the observed-score scale of the test. The transformation is amply demonstrated in Lord and Novick's (1968, sect. 16.14) well-known graphs of "typical distortions in mental measurement." Tests with identical characteristic functions produce the same true-score distributions if the ability distribution of the examinees is the same. For professional tests of sufficient length, with items produced by trained item writers, the reliability coefficients typically are in the upper .80s or lower .90s. Therefore, differences between the shapes of the observed-score and true-score distributions are usually minor compared to the differences between the observed-score distribution and the ability distribution on the Θ-scale. Also, a target for the characteristic function of the test implicitly constrains the information function of the test to have its larger values in the region where the characteristic function has its steepest slope. Typically, the ability distribution is centered in this region, and therefore the impact of random error on the true-score distribution for

the test is automatically reduced for the majority of the examines. However, it is a straightforward extension to provide the model with explicit constraints for the information function of a test.

An attractive feature of the test characteristic function is that, like the test information function, it is additive across the items. This facts allows us to design Linear Programming (LP) models for test assembly that minimize the differences between a test characteristic function and its target. LP models for test assembly have been introduced earlier for a variety of other test assembly problems (Adema, 1990a, 1990b, 1992; Adema & van der Linden, 1989; Amstrong & Jones, 1992; Amstrong, Jones & Wu, 1992; Boekkooi-Timminga, 1987, 1989, 1990a, 1990b; Theunissen, 1985; van der Linden, 1993; van der Linden & Boekkooi-Timminga, 1988, 1989).

Model

The following notation is needed to present the model. Let i=1,...,I denote the items in the pool and let $x_i=\{0,1\}$ be decision variables to denote whether or not the item will be assigned to the test. Suppose that the test characteristic function, which is defined as the sum of the item response functions $P_i(\Theta)$ in the test, has to be controlled for a grid of fixed ability values Θ_k, k=1,...,K. The target values for the test characteristic function are denoted by $T_C(\Theta_k)$. Finally, the positive and negative deviations of the test characteristic function from its target values are defined as (non-negative) variables u_k and v_k, respectively. Then the following model minimizes the sum of the deviations of the test characteristic function form its target values:

$$minimize \quad \sum_{k=1}^{K} (u_k+v_k) \tag{1}$$

subject to

$$\sum_{i=1}^{I} P_i(\theta_k)x_i-u_k+v_k = T_C(\theta_k), \qquad k=1,...,K; \tag{2}$$

$$\sum_{i=1}^{I} x_i = n; \tag{3}$$

$$\sum_{i\varepsilon V_j} x_i \geq n_j^{(1)}, \qquad j=1,...,J; \tag{4}$$

$$\sum_{i\varepsilon V_j} x_i \leq n_j^{(2)}, \qquad j=1,...,J; \tag{5}$$

$$x_i = 0,1, \qquad i=1,...,I; \tag{6}$$

$$u_k,\ v_k \geq 0.$$

In (2) the variables u_k and v_k are defined. The constraint in (3) puts the length of the test equal to n items. The constraints in (4) and (5) impose lower and upper bounds to the numbers of items to be selected from subsets V_j, j=1,...,J, in the item pool, where each subset V_j is supposed to cover a content area represented in the pool. These constraints will be used in the example below to guarantee that existing content specifications for the test are met.

The constraints in the model are a small sample of the possibilities available to realize test specifications when assembling tests through the use of LP models. Any specification that can be represented as a linear (in)equality in the decision variables can be inserted in the model. A review of other possibilities is given in van der Linden and Boekkooi-Timminga (1989). Algorithms and heuristics for solving LP models for test assembly are described in Adema, Boekkooi-Timminga and van der Linden (1991), Amstrong, Jones and Wu (1992) and Luecht and Hirsch (1992).

An Empirical Example

To illustrate the practical use of the model in this paper, a test was assembled from an item pool previously in use for the Mathematics Test in the ACT Assessment Program (AAP). The pool consisted of 520 items all calibrated under the 3-parameter logistic model using an MML method with Θ distributed as N(0,1).

Method

The following steps were taken in this study:

First, a 40-item test, assembled by hand to meet the specifications in the AAP at an earlier occasion, was selected from the pool to generate a target for the distribution of the observed scores. The target was generated assuming the abilities in the population of examinees to be distributed N(0,1) and using the generalized binomial as the conditional probability function of the observed score given the ability level of the examinee (Lord, 1980, sect. 4.1).

Second, the relative true-score distribution associated with the observed-score distribution was assumed to follow a four-parameter beta density with function:

$$g(\tau)=n^{-1}(-l+\tau)^{a-1}(u-\tau)^{b-1}/(u-l)^{a+b-1}B(a,b), \qquad (8)$$

where τ is the relative true score, B(a,b) is the Beta function with parameters a and b, and the density is defined on the interval [l,u] with $0\leq l<u\leq 1$. All four unknown parameters were estimated from the first four factorial moments of the target for the observed-score distribution using a program by Hanson (1991).

Three, because the test characteristic function transforms the Θ-scale into the true-score scale, it can be calculated from the distribution functions of the abilities and the true scores. Let $G(\tau)$ be the distribution function associated with the beta density in (8) and $F(\Theta)$ the N(0,1) distribution function. Then the test characteristic function is given by:

$$T_C(\theta) = 40G^{-1}(F(\theta)). \quad (9)$$

Four, target values for the test characteristic function were calculated from (9) and inserted into the constraint in (2). The model was solved to assemble a 40-item test from the pool with a test characteristic function meeting the target values in (2). The model was solved using an adapted version of the heuristic in Luecht and Hirsch (1992).

Five, two different versions of the model were solved. One model was the full model with the content constraints in (4)-(5). The following six content areas were represented in the pool: Arithmetic and Algebraic Reasoning (14); Arithmetic and Algebraic Operations (4); Geometry (8); Intermediate Algebra (8); Number and Numeration Concepts (4); and Advanced Topics (2). The numbers between parentheses are the required numbers of items in the test for each of the content areas. The second model ignored all content constraints.

Six, for both solutions the observed-score distributions were generated using the same procedure as in Step 1.

Results

The characteristic functions of the tests assembled without and with the content constraints are presented in Figures 1 and 2, respectively. Each functions appears to closely approximate its respective target characteristic function. The effects of imposing content constraints on the assembly process seem to be negligible. Figure 3 plots the difference between the test characteristic functions in Figures 1-2 as a function of Θ. The difference is never larger than .26 on the true-

score scale, which runs from 0-40, whereas the mean difference is equal to .18.

Figures 1-3 about here

In Figures 4-5 the observed-score distributions generated for the two solutions are plotted. Both for the model with and the model without the content constraints the distributions fit the distribution of the original target test tightly over the whole score range, except for a small bump just to the left of the middle of the scale. It is unclear to the authors whether these bumps, which were a systematic phenomenon in runs with other problems by the authors, are caused by the actual composition of the item pool and/or features of the heuristic used to solve the model. As displayed in Figure 6, the mean difference between the two distributions is equal to zero and is never larger in absolute value than .0008 across the observed-score scale.

Figures 4-6 about here

Discussion

The empirical study should be repeated for other item pools and test assembly problems to provide further support for the practical feasibility of the model presented in this chapter. Also, it might be worthwhile to study the effect of introducing a target for the test information function as an additional contraint in the

model. Such a target could be used for fine tuning the observed-score distribution in certain regions, for instance, at its right-hand tail if the test is used to award scholarships to the best students.

The remarkable thing about the method followed in the empirical example is that no distribution of *actual* observed scores is required to set a target for the test; the only information needed is the density of this distribution. In principle, all a test assembler has to do is to draw a curve on paper that represents the density of the observed-score distribution he or she has in mind. The method of moments, commonly in use as a method for estimating the parameters in the beta-binomial model and implemented in the program by Hanson used in the empirical example in this paper, allows us to estimate the target for the true-score distribution directly from this curve, and from there on it is only one step to derive a target for the characteristic function of the test. However, in addition to this approach, it is always possible to administer a real test to a random sample of examinees for the population for which the test program is designed, and use its scores as a target for the observed-score distribution in the program.

References

Adema, J.J. (1990a). The construction of customized two-staged tests. Journal of Educational Measurement, 27, 241-253.

Adema, J.J. (1990b). Models and algorithms for the construction of achievement tests. Ph.D. thesis, University of Twente, Enschede, The Netherlands.

Adema, J.J. (1992). Methods and models for the construction of weakly parallel tests. Applied Psychological Measurement, 16, 53-63.

Adema, J.J., Boekkooi-Timminga, E., & van der Linden, W.J. (1991). Achievement test construction using 0-1 linear programming. European Journal of Operations Research, 55, 103-111.

Adema, J.J. & van der Linden, W.J. (1989). Algorithms for computerized test construction using classical item parameters. Journal of Educational Statistics, 14, 279-290.

Amstrong, R.D. and Jones, D.H. (1992). Polynomial algorithms for item matching. Applied Psychological Measurement, 16, 365-373.

Amstrong, R.D., Jones, D.H., & Wu, I-L. (1992). An automated test development of parallel tests. Psychometrika, 57, 271-288.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord and M.R. Novick (1968), Statistical theories of mental test scores. Reading, Massachusetts: Addison-Wesley.

Boekkooi-Timminga, E. (1987). Simultaneous test construction by zero-one programming. Methodika, 1, 1101-112.

Boekkooi-Timminga, E. (1989). Models for computerized test construction. Ph.D. thesis, University of Twente, Enschede, The Netherlands.

Boekkooi-Timminga, E. (1990a). The construction of parallel tests from IRT-based item banks. Journal of Educational Statistics, 15, 129-145.

Boekkooi-Timminga, E. (1990b). A cluster-based method for test construction. Applied Psychological Measurement, 15, 129-145.

Gulliksen, H. (1950). Theory of mental tests. New York: Wiley.

Hanson, B.A. (1991). Method of moments estimates for the four-parameter compound binomial model and the calculation of classification consistency indices (ACT Research Report Series 91-5). Iowa City, IA: American College Testing.

Lord, F.M. (1965). A strong true-score theory, with applications. Psychometrika, 30, 239-270.

Lord, F.M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum.

Lord, F.M. & Novick, M.R. (1868). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Luecht, R.M. & Hirsch, T.M. (1992). Computerized test construction using average growth approximation of target information functions. Applied Psychological Measurement, 16, 41-52.

Samejima, F. (1977). Weakly parallel tests in latent trait theory with some criticism of classical test theory. Psychometrika, 42, 193-198.

Theunissen, T.J.J.M. (1985). Binary programming and test design. Psychometrika, 50, 411-420.

van der Linden, W.J. (1993). Optimum design in item response theory: Applications to test assembly and item calibration. in G.H. Fischer & D. Laming (Eds,), Contributions to mathematical psychology, psychometrics, and methodology (pp. 303-316). New York: Springer-Verlag.

van der Linden, W.J. & Boekkooi-Timminga, E. (1988). Á zero-one programming approach to Gulliksen's matched random subsets method. Applied Psychological Measurement, 12, 201-209.

van der Linden, W.J. & Boekkooi-Timminga, E. (1989). A maximin model for test design with practical constraints. Psychometrika, 53, 237-247.

Authors' Note

The authors are indebted to Peter Yang for his computational assistance in preparing the empirical example in this paper. This paper was written while both authors were at American College Testing, Iowa City, IA, USA. Richard Luecht is now at the National Board of Medical Examinees, Philadelphia, PA, USA.

Figure Captions

Figure 1. Comparison between the characteristic function of the assembled test and its target (model without content constraints)

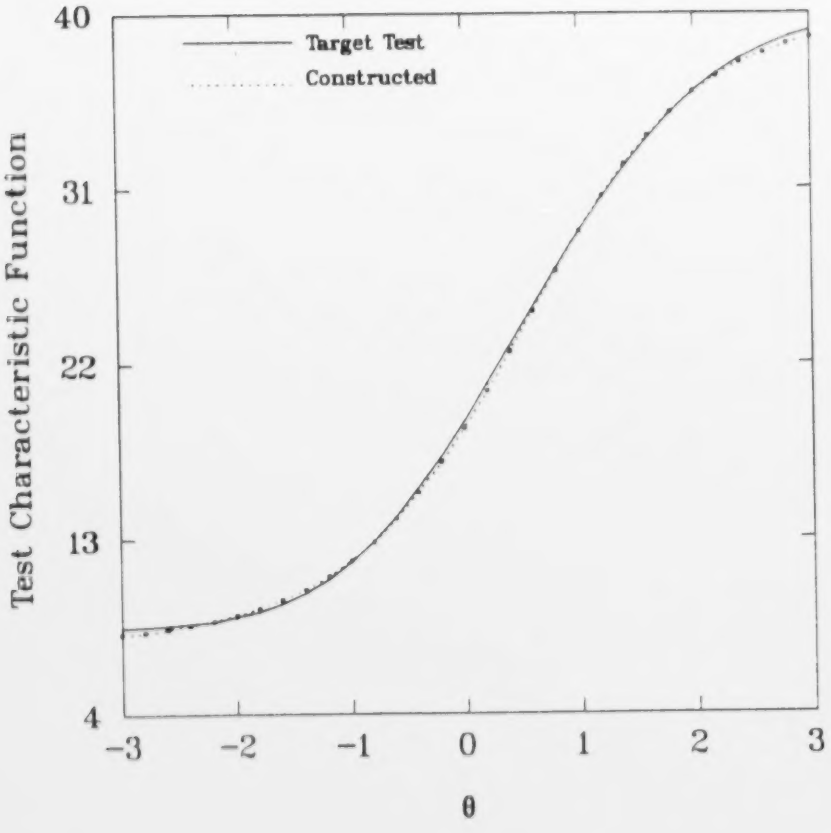
Figure 2 Comparison between the characteristic function of the assembled test and its target (model with content constraints)

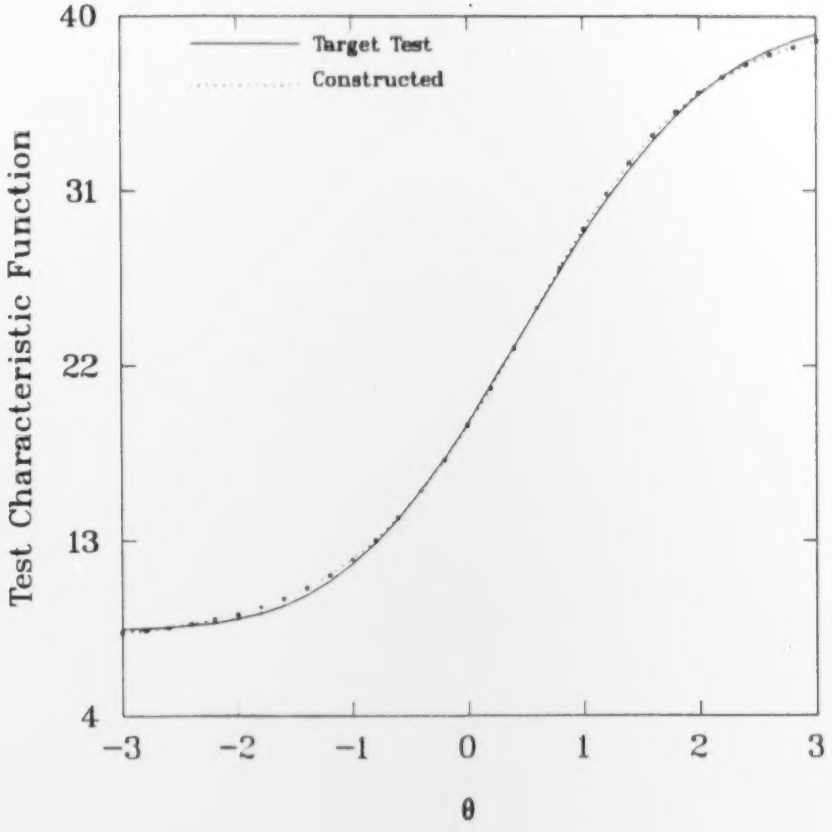
Figure 3 Differences between the characteristic functions of the assembled tests in Figures 1-2 as a function of Θ.

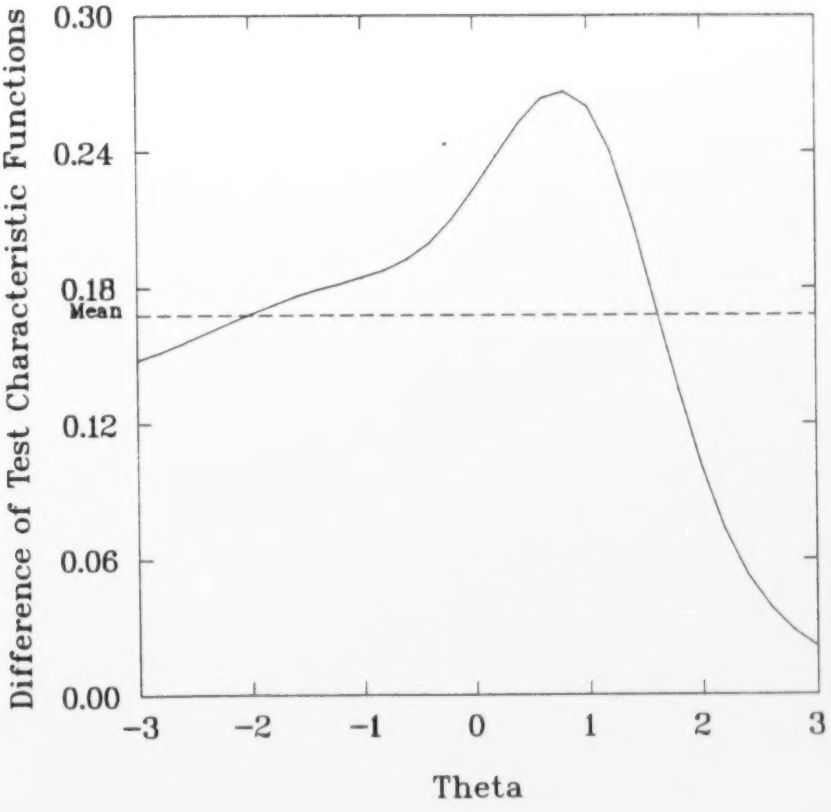
Figure 4 Comparison between the density function of the observed-score distribution on the assembled test and its target (model without content constraints)

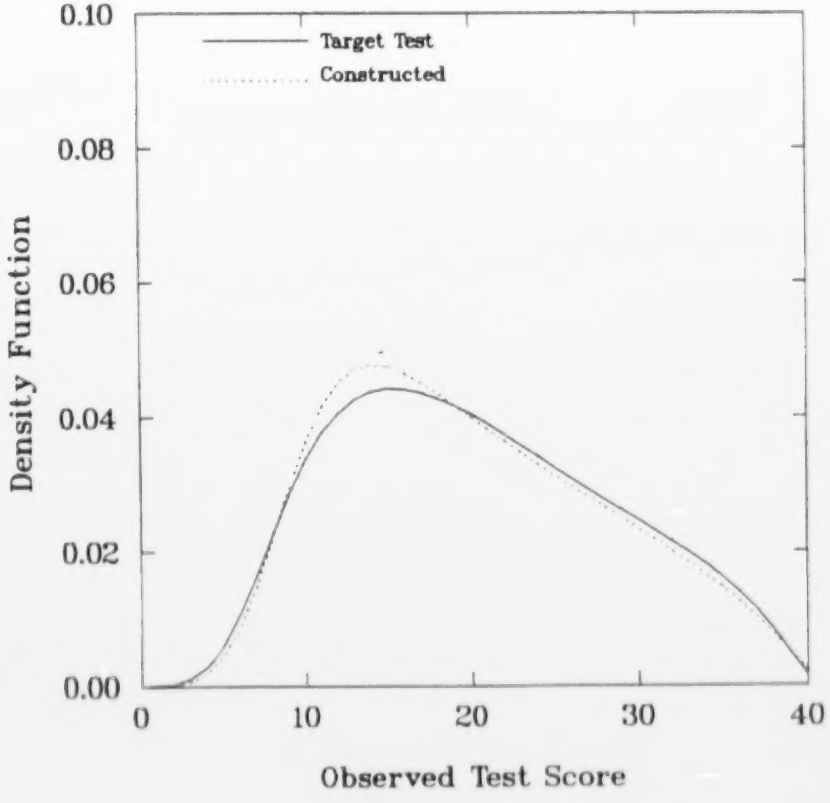
Figure 5 Comparison between the density function of the observed-score distribution on the assembled test and its target (model with content constraints)

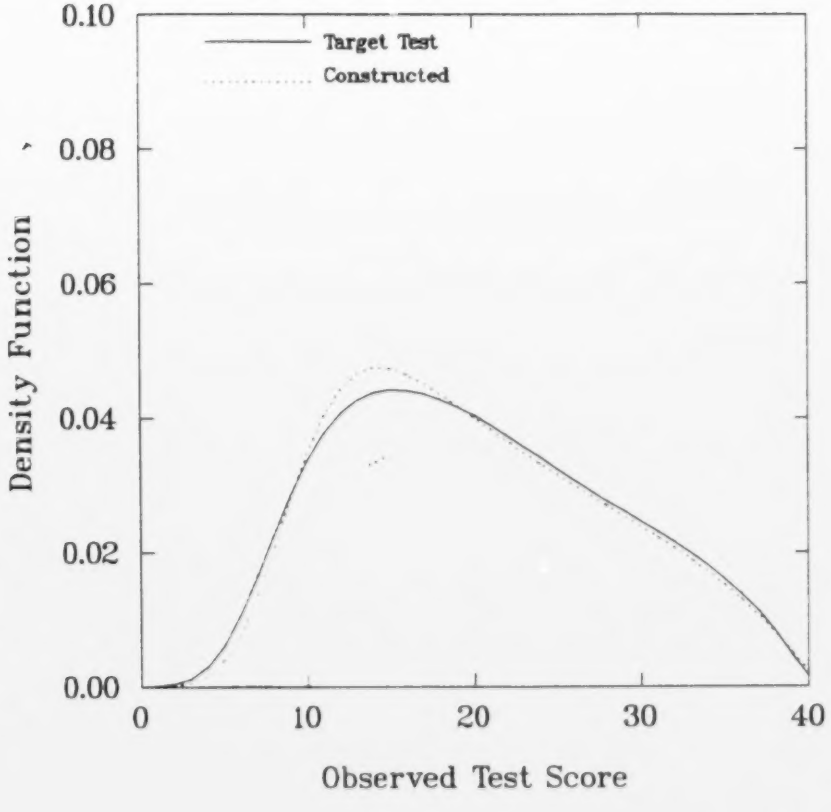
Figure 6 Differences between the density functions of the assembled tests in Figures 4-5 as a function of Θ.

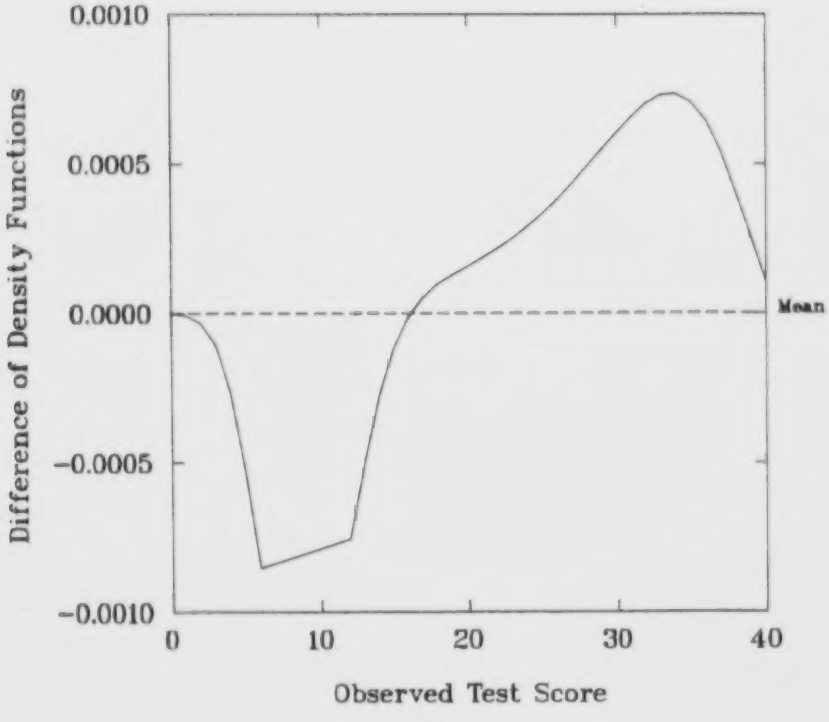

Target Test
Constructed
Test Characteristic Function
40
31
22
13
4
−3
−2
−1
0
1
2
3
θ


Target Test
Constructed
Test Characteristic Function
40
31
22
13
4
−3
−2
−1
0
1
2
3
θ


Difference of Test Characteristic Functions
0.30
0.24
0.18
Mean
0.12
0.06
0.00
−3
−2
−1
0
1
2
3
Theta


Target Test
Constructed
Density Function
0.10
0.08
0.06
0.04
0.02
0.00
0
10
20
30
40
Observed Test Score


Target Test
Constructed
Density Function
0.10
0.08
0.06
0.04
0.02
0.00
0
10
20
30
40
Observed Test Score


0.0010
0.0005
0.0000
−0.0005
−0.0010
Difference of Density Functions
Mean
0
10
20
30
40
Observed Test Score

Titles of recent Research Reports from the Department of Educational Measurement and Data Analysis. University of Twente, Enschede, The Netherlands.

RR-94-7 W.J. van der Linden & R.M. Luecht, *An optimization model for test assembly to match observed-score distributions*

RR-94-6 W.J.J. Veerkamp & M.P.F. Berger, *Some new item selection criteria for adaptive testing*

RR-94-5 R.R. Meijer, K. Sijtsma & I.W. Molenaar, *Reliability estimation for single dichotomous items*

RR-94-4 M.P.F. Berger & W.J.J. Veerkamp, *A review of selection methods for optimal design*

RR-94-3 W.J. van der Linden, *A conceptual analysis of standard setting in large-scale assessments*

RR-94-2 W.J. van der Linden & H.J. Vos, *A compensatory approach to optimal selection with mastery scores*

RR-94-1 R.R. Meijer, *The influence of the presence of deviant item score patterns on the power of a person-fit statistic*

RR-93-1 P. Westers & H. Kelderman, *Generalizations of the Solution-Error Response-Error Model*

RR-91-1 H. Kelderman, *Computing Maximum Likelihood Estimates of Loglinear Models from Marginal Sums with Special Attention to Loglinear Item Response Theory*

RR-90-8 M.P.F. Berger & D.L. Knol, *On the Assessment of Dimensionality in Multidimensional Item Response Theory Models*

RR-90-7 E. Boekkooi-Timminga, *A Method for Designing IRT-based Item Banks*

RR-90-6 J.J. Adema, *The Construction of Weakly Parallel Tests by Mathematical Programming*

RR-90-5 J.J. Adema, *A Revised Simplex Method for Test Construction Problems*

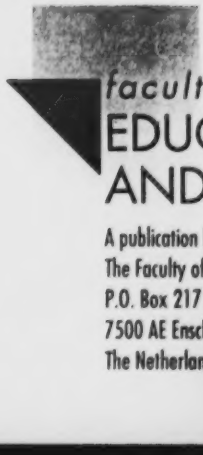
RR-90-4 J.J. Adema, *Methods and Models for the Construction of Weakly Parallel Tests*

RR-90-2 H. Tobi, *Item Response Theory at subject- and group-level*

RR-90-1 P. Westers & H. Kelderman, *Differential item functioning in multiple choice items*

Research Reports can be obtained at costs from Bibliotheek, Faculty of Educational Science and Technology, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.

BEST COPY AVAILABLE



faculty of
EDUCATIONAL SCIENCE AND TECHNOLOGY

A publication by
The Faculty of Educational Science and Technology of the University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

BEST COPY AVAILABLE